On Selecting Online Abbreviation Dictionary – Technical Report

Maciej Gawinecki

ICT School, University of Modena and Reggio Emilia,
Via Vignolese 905, 41100 Modena, Italy
Maciej.Gawinecki@unimore.it

Abstract. Schema element names often contain abbreviations which need to be expanded first before lexical annotation is performed on them. When neither the context of abbreviation occurrence nor the corresponding documentation can provide a corresponding expansion, then an online abbreviation dictionary remains "last chance" source for automated abbreviation expansion. Such a solution is even more general in comparison to user-defined dictionary as it does not requires a user to maintain it. It in this work we report a short survey on selecting the best online abbreviation dictionary and describe how it can be easily integrated with abbreviation expansion component in MOMIS schema integration system.

1 Introduction

Schema element names often contain abbreviations that need to be expanded first before lexical annotation is performed on them. When neither the context of abbreviation occurrence nor the corresponding documentation can provide a corresponding expansion, then an online abbreviation dictionary remains the "last chance" source for automated abbreviation expansion. The purpose of an online abbreviation dictionary is to provide comprehensive high-quality coverage of abbreviation long forms while still keeping the collection current as the new abbreviations continually come into use. This way using abbreviation expansion is even more general solution in comparison to user-defined dictionary as it does not require a user to maintain it.

Online abbreviation dictionaries have been successfully used as source of abbreviation long forms in the context of cleaning text from on-line chats [1]. However, upon our knowledge there is no schema matching nor annotation tool that use them. As there are several abbreviation dictionaries available online, we conducted a short survey (Section 2) on them to pick up one, that the best fits for the problem of abbreviation expansion in schema element names. Once selected, the dictionary must be integrated with the schema matching component. In Section 3 we addressed such integration issues in the context of MOMIS data integration system [2, 3].

For clear understanding, we use the term *short form* to refer to an abbreviation, and *long form* for its corresponding full word expansion.

2 Survey on online abbreviation dictionaries

There are several abbreviation dictionaries online. They differ in: language of long forms, domain coverage (general or domain-specific, as for instance *Dog fanciers acronym list*[1] or *Mad Cow disease list*[2]), up-to-date coverage (whether it keeps the collection current as the new abbreviations continually come). In Tables 1, 2 and 3 we present a comparison of a subset of them, the most accepted ones in the community. We would like to choose the one that satisfy the following requirements:

- *English-specific* – a dictionary contains English abbreviations,
- *wide coverage* – a dictionary is not domain specific but rather covers, at least superficially, all domains,
- *high quality of data* – definitions are either (i) proposed manually and thoroughly verified or (ii) collected automatically with proven threshold of relevance (precision with recall),
- *disambiguation information* – as there are can be more then one long form for a given short form, an information for disambiguating long form candidates is required; this includes ranking (in respect to accuracy, relevance, popularity or commonness) of long forms and knowledge domain to which the long form belongs or category, in which the long form is used
- *language information* – provide some mean for language disambiguation among different long forms of a single short form to prune non-English long forms
- *publicly accessible* – it must be possible either to (i) *download* a dictionary and use it offline or (ii) access the dictionary *remotely* (preferably via Web service API) and cache information on call; obviously, permission given by a dictionary license must allow for such an access for research purposes.

2.1 Manually constructed dictionaries

In Tables 1 and 2 we have compared 6 manually constructed dictionaries covering general domain of interest: *Abbreviations.com*[3], *Sigles.net*[4], *The Internet Acronym Servers*[5], *Acronym Finder*[6], *WordNet*[7] and *All-acronyms.com*[8]. It is necessary to point out here, that one of them, Acronym Finder, was also partially contributed by results of automatic extraction from Acrophile project [4, 5]. We decided to include also WordNet as an abbreviation dictionary, because

[1] `http://mx.nsu.ru/FAQ/F-dogsacronym-list/Q0-0.html`
[2] `http://www.maff.gov.uk/animalh/bse/glossary.html`
[3] `http://abbreviations.com`
[4] `http://www.sigles.net`
[5] `http://silmaril.ie/cgi-bin/uncgi/acronyms`
[6] `http://www.acronymfinder.com`
[7] `http://wordnet.princeton.edu`
[8] `http://www.all-acronyms.com`

it also contains expansions for some abbreviations. For instance, the abbreviation *sec* is in WordNet defined as a shortcut for *second*, *secant* and *Security and Exchange Commission* (SEC). We did not include popular *The Free Dictionary By Farlex*[9], because it is powered by already included the *Acronym Finder* dictionary. We decided also not to compare number of long forms vs. number of short forms provided by particular dictionaries, because information provided on their sites is often either incomplete (only number of long forms or number of only human-defined forms is given) or unclear (e.g. what number of abbreviation means – number of long forms or short forms).

Access to a dictionary. With regard to access to dictionary content, only WordNet permit to download the whole content. This is not a surprise, as it is the only site in this competition, that is a pure academic project; the rest of dictionaries are commercial initiatives (for instance All-acronyms.com is a site promoting works of a publisher of abbreviation dictionaries). On the other hand, currently, only Abbreviations.com dictionary provides Web service access to its content, making it easier to integrate the dictionary into our solution. WWW access does not exclude a site from the usage, by writing a wrapper over a HTML document source makes this task more tedious. For instances, authors of [1] have used Perl scripts with regular expressions to scrape and cache data from Abbreviations.com [6].

Information for disambiguation of long forms. Four dictionaries provide ranking of long forms for a single abbreviation. The ranking is constructed in respect to either definition's popularity, commonness, relevance or accuracy, thought it is not clear how this factors has been measured or at least estimated in any of the analyzed dictionaries. Five from the six analyzed dictionaries provided information about corresponding definition domain. However, this information has been differently structured in different abbreviation dictionaries. Two of them, Abbreviations.com and WordNet, provides a hierarchy of categories/domains and a long form is annotated with at least one of them (hierarchical multiply domain information). Moreover, the Abbreviations.com dictionary provides a popularity information of a long form in each domain separately. For instance, abbreviation Id is in the dictionary expanded to (top 5 positions): identification (*Governmental* ▷ *Military*), identification (*Medical* ▷ *Physiology*), identification (*Governmental* ▷ *Police*), Indonesia (*Regional* ▷ *Countries*), identifier (*Computing* ▷ *Drivers*). Other dictionaries has a simple set of categories, and a definition can be annotated either with a single category (flat domain information), as in Acronym Finder or a combination of categories (multiply flat domain information), as in Sigles.net, All-acronyms.com.

Language of a long form. Three dictionaries are known to provide definitions in languages other than English (about the rest three dictionaries no information has been found). In such context we were interested how these three dictionaries allows to recognize a language of definition, so we could filter out non-English long forms. Sigles.net provide additional annotation about a language of the definition. Non-English long forms in Abbreviations.com do not

[9] `http://acronyms.thefreedictionary.com`

have defined domain category, but instead are categorized into one of the 'International' subcategories. In this way no domain information can be obtained for non-English long forms, but fortunately we are interested only in English long forms. Acronym Finder provides a less convenient way to distinguish language of a long form, by providing information about it as a part of long form string, e.g. *"Unit d'Enseignement (French: academic subjects)"*.

2.2 Automatically constructed dictionaries

The purpose of an online abbreviation dictionary is to provide comprehensive high-quality coverage of abbreviation long forms while still keeping the collection current as the new abbreviations continually come into use. Satisfying these requirements is a challenging task, which has been solved in manually constructed dictionaries by allowing users to submit new abbreviations and by carefully manually verifying their suggestions.

However, this requires a lot of human effort and can still result in not so high quality of data, especially in technical fields which are rapidly changing, like medicine or computer technologies. In this cases using a manually created abbreviation dictionary is not possible, because [7]: (1) acronyms and abbreviations may be created faster than they can be added to such a dictionary, (2) specialized dictionaries for particular sub-fields are not always available. This problem has been addressed by automated building abbreviation dictionaries [4, 8, 7]. In this case abbreviations having likely definitions are identified and the long forms extracted from the documents corpus using techniques based on abbreviation patterns discovery [9, 10, 4, 8, 11]. The definitions having identical associated abbreviations are grouped together, then definition groups are arranged into clusters to determine a similar definition. Further disambiguation can be provided by looking at similarity between clusters using an annotation associated with the source documents from which the definitions where extracted [7].

In Table 3 we have compared 2 automatically constructed dictionaries: Acrophile[10] and Biomedical Abbreviation Server[11]. The main drawbacks of Acrophile that excludes it from usage in our case are is a need to provide it with new corpus of sites to update abbreviations database, lack of direct domain information, and lack of Web service access[12]. On the other side Biomedical Abbreviation Server covers only medical domain.

2.3 Selection of dictionary

None of dictionary is ideal for us, but we have chosen Abbreviations.com, because it is very popular in the domain and has been recognized in the community with many prizes. Moreover, although it is not possible to download all long

[10] `http://ciir.cs.umass.edu/irdemo/acronym/index.html`
[11] `http://abbreviation.stanford.edu/`
[12] For detailed results of evaluation and comparison of Acrophile with manually created dictionary see [4].

forms from the site, it can be easily integrated with any annotation application, because it provides Web service API (though it can be limited by number of requests/day). It also has a mean to distinguish English long forms and provides long form domain and popularity of long forms, that we will valuable during disambiguation phase.

3 Integration issues

3.1 Mapping categories from Abbreviations.com to WordNet Domains

One of possible disambiguation methods is relating the relevance of a long form candidate to the number of domains shared with an annotated schema. However, this can be not a straightforward task, because MOMIS annotation algorithm uses WordNet Domains[13] to define prevalent domains of the schema [3], while Abbreviations.com employs its own long forms classification system. Therefore a mapping between both need to be defined.

Methodology of mapping discovery. Obviously, the mapping does not need to be bidirectional. It rather has to answer to the question: to which WordNet domains may belong abbreviations from the given Abbreviations.com category? Unfortunately, this is not so trivial tasks, because each categorization system has been designed for slightly different application purpose. Abbreviations.com categories has been selected to classify context in which short forms and their corresponding long forms appear nowadays. More precisely, very often they signify not what the long form expresses (e.g. that HB – Happy Birthday does not relate to time period), but when, where or by what community it is used (HB is shortcut used in SMS messages). This way it includes such categories as NASDAQ symbols, HTTP acronyms etc. On the other hand WordNet Domains 3 has been designed to support computational linguistics (including word sense disambiguations, text categorization and information retrieval) and is based on Dewey Decimal Classification (DDC), the system used by librarians to classify books in respect to what part of knowledge they describe [12]. The name of each each domain represents a discipline where a certain knowledge area is developed, e.g. chemistry or a specific object of the knowledge area, e.g. food. To realize a task of finding a mapping between this two categorization systems we need to know the precise semantics of each category/domain in both systems.

The semantics of each domain in WordNet Domains hierarchy is determined by [12]:

- a short lexical description of the domain
- hierarchical relations with other domains,
- synsets in WordNet which belongs to the given domain and
- DDC classes to which the domain has been assigned.

The semantics of each category in Abbreviations.com can be determined by:

[13] `http://wndomains.itc.it/wordnetdomains.html`

– again, a short lexical description of the category,
– hierarchical relations with other categories and
– abbreviations which belongs to the category.

We automatically created initial mapping, following the intuition that a category can be mapped to one or more WordNet domains, if its name is equal to the name of a domain or is a synset belong to such a domain. For multi-word category names we summed up domains of constituting word, while if no domain was found at for for a particular word we repeated our trial for the base form of the word (using Porter stemmer [13]). Such initial mapping has shown that the same names in both categorization systems can cover not exactly the same part of the world or no corresponding domain can be found for a particular category name (because there is no lemma in WordNet with such a name). Therefore, we carefully revised our initial mapping on the base of semantics of abbreviations belonging to questionable categories.

Results of mapping. The complete mapping can be found at: `http://www.mapping.it`. For many categories that contain abbreviations that would be never used in schema element names or are too generic we decided to assign *Factotum* domain. This is the case for such categories as: *Wannas* (Words written with American or Afro-American slang, like 'KINDA' – 'Kind Of' – or 'MZ' – ' slang for MRS'), *MIME* (Multipurpose Internet Mail Extensions, like .CLASS), *Conferences* (meetings of people that "confer" about a particular topic, e.g. business, sport, science etc.) or *Famous* (e.g. '007' – James Bond, 'JPII' – 'Pope John Paul II'). In this way we assume long form candidates, that are less relevant or can be confusing for results will not be promoted. This is because Factotum domain is not used by domain disambiguation algorithms [14, 3].

There is also one special category *Journal Abbreviation Sources*, which has not been considered at all, because it is rather "a registry of Web resources that list or provide access to the full title of journal abbreviations".

3.2 Estimating popularity

Popularity of long form candidates is not reported explicitly by the dictionary Web service API but can be easily estimated from the order in which dictionary entries are reported.

3.3 Ambiguous categories reported

When returning corresponding long form, the service reports them together with only names of sub-categories, to which they belong, without names of their parental top-categories. Since there are two sub-categories, one in 'Computing' top-category, and another one in 'Business' top-category, both identified by the same name 'General' – such form of reporting can lead to ambiguous interpretation.

3.4 Message-level integration

Abbreviations.com provides access to its content via RESTFul Web service API[14] under the following URL: `http://www.abbreviations.com/services/v1/abbr.aspx`. The complete description of request parameters can be found in Table 4. A response to a request is a XML document in the self-explaining form, described in Table 5.

For instance, let us assume we are looking in the Abbreviation.com for longs forms of 'id' short form. The request would be the following URL (combination of base service URL and a query string):

```
http://www.abbreviations.com/services/v1/abbr.aspx
  ?tokenid=tk324324324&term=id
```

The sample response to this request would be:

```
<?xml version="1.0" encoding="UTF-8"?>
<results>
  <result>
    <term>id</term>
    <definition>Identification</definition>
    <category>Military</category>
  </result>
  <result>
    <term>id</term>
    <definition>Identification</definition>
    <category>Physiology</category>
  </result>
  <result>
    <term>id</term>
    <definition>Indonesia</definition>
    <category>Country</category>
  </result>
</results>
```

On the implementation level we are using Jersey, JAX-RS (JSR 311) Reference Implementation from Sun for building RESTful Web services[15] to communicate and JAXB Reference Implementation from Sun for Java Architecture for XML Binding[16] to access XML message content from Java level. For the latter implementation it is important to notice that Abbreviations.com does not explicitly define XML schema of messages and thus a developer of a service client is responsible for foreseeing and handling possible message content, including error reports, e.g.:

[14] RESTful Wen services, `http://en.wikipedia.org/wiki/Representational_State_Transfer`.

[15] `https://jersey.dev.java.net/`

[16] `https://jaxb.dev.java.net/`

```
<?xml version="1.0" encoding="UTF-8"?>
<results>Invalid Token</results>
```

4 Conclusions and Future Work

In this short report we conducted a survey on selecting online abbreviation dictionary, that would be useful for abbreviation expansion in context of schema matching. We also defined a mapping between selected dictionary categorization system and WordNet Domains used by MOMIS data integration system. Our future experiments will show what is the quality of the created mapping and its impact on the effectiveness of planned abbreviation expansion techniques, using this mapping.

5 Acknowledgments

Many thanks go to Younchahou Fifen Njikam for the discussion on the online abbreviation dictionaries and recognizing abbreviation/definition patterns in texts.

References

1. Wong, W., Liu, W., Bennamoun, M.: Integrated scoring for spelling error correction, abbreviation expansion and case restoration in dirty text. In: AusDM '06: Proceedings of the fifth Australasian conference on Data mining and analystics, Darlinghurst, Australia, Australia, Australian Computer Society, Inc. (2006) 83–89
2. Bergamaschi, S., Castano, S., Vincini, M.: Semantic integration of semistructured and structured data sources. SIGMOD Rec. **28**(1) (1999) 54–59
3. Bergamaschi, S., Po, L., Sorrentino, S.: Automatic annotation for mapping discovery in data integration systems. In: Sixteenth Italian Symposium on Advanced Database Systems (SEBD 2008). (2008) 334–341
4. Larkey, L.S., Ogilvie, P., Price, M.A., Tamilio, B.: Acrophile: an automated acronym extractor and server. In: DL '00: Proceedings of the fifth ACM conference on Digital libraries, New York, NY, USA, ACM (2000) 205–214
5. Ogilvie, P. personal communication (12 2008)
6. Wong, W. personal communication (3 2009)
7. Adar, E., Adamic, L.A.: Method and system for building an abbreviation dictionary (14 2006)
8. Chang, J.T., Schtze, H., Altman, R.B.: Creating an online dictionary of abbreviations from MEDLINE. Journal of the American Medical Information Association **9**(6) (2002) 612–620
9. Yeates, S.: Automatic Extraction of Acronyms from Text. In: Proceedings of the Third New Zealand Computer Science Research Students Conference, Hamilton, New Zealand, University of Waikato (April 1999) 117–124
10. Yeates, S., Bainbridge, D., Witten, I.H.: Using Compression to Identify Acronyms in Text. In: DCC '00: Proceedings of the Conference on Data Compression, Washington, DC, USA, IEEE Computer Society (2000)

11. Hill, E., Fry, Z.P., Boyd, H., Sridhara, G., Novikova, Y., Pollock, L., Vijay-Shanker, K.: AMAP: automatically mining abbreviation expansions in programs to enhance software maintenance tools. In: MSR '08: Proceedings of the 2008 international working conference on Mining software repositories, New York, NY, USA, ACM (2008) 79–88
12. Bentivogli, L., Forner, P., Mangini, B., Pianta, E.: Revising the Wordnet Domains Hierarchy: semantics, coverage and balancing. In Sérasset, G., ed.: COLING 2004 Multilingual Linguistic Resources, Geneva, Switzerland, COLING (August 28 2004) 94–101
13. Porter, M.F.: An algorithm for suffix stripping. Program **14**(3) (1980) 130–137
14. Buscaldi, D., Rosso, P., Masulli, F.: Integrating Conceptual Density with WordNet Domains and CALD Glosses for Noun Sense Disambiguation. In: EsTAL. (2004) 183–194

Table 1. Comparison of manually created on-line abbreviation dictionaries.

Dictionary	*Abbreviations.com*	*Sigles.net*	*The Internet Acronym Servers*	*Acronym Finder*	*WordNet*	*All-acronyms.com*
Year of creation:	2001	2005	1988	1998	1985	2005
Domain coverage:	general	general	general	general	general	general
How it was built:	manually	manually	manually	manually with some contribution of automatic extraction	manually	manually
Submission and verification policy:	via registered volunteer editors	via unregistered volunteer editors and verified by main editor	via unregistered volunteer editors and verified by main editor	via unregistered volunteer editors and verified by main editor	in terms of research	via unregistered volunteer editors and verified by main editor
Access:	WWW, REST Web service	WWW	WWW, Web service (planned)	WWW	WWW, download	WWW
Ranking of long forms:	based on popularity	no	no	subjective measure of commonness, popularity or relevance	based on popularity	probably based on accuracy and popularity
Domain information:	multiply hierarchical	multiply flat	no	flat	multiply hierarchical	multiply flat

Table 2. Comparison of manually created on-line abbreviation dictionaries (continuated).

Dictionary	*Abbreviations.com*	*Sigles.net*	*The Internet Acronym Servers*	*Acronym Finder*	*WordNet*	*All-acronyms.com*
Language of long forms:	English, Spanish, French, Mexican, Russian, Italian, Latin, German, Turkish, Hebrew	35 languages, including: French, English (UK, USA, Canada, Australia), Spanish, Italian, Marocinian	uknown	English (UK/USA, Australian), French, Spanish, Portuguese, Italian, Pakistani, Polish, German and others	English	mostly English (99.8%)
Language information:	via annotation with one of *International* subcategories	yes	no	a language is contained in definition's name, e.g. *"Unit d'Enseignement (French: academic subjects)"*	-	-
License:	Web service API is limited to 1,000 queries for a developer token per day	only for education usage	unknown	automated extracting data from a web page is strongly prohibited	free for research purposes	free for research purposes

Table 3. Comparison of automatically created on-line abbreviation dictionaries.

Dictionary	*Acrophile*	*Biomedical Abbreviation Server*
Year of creation:	2000	2002
Domain coverage:	general	bio-medical
Number of short forms:	51,726	2,074,367
Number of long forms:	161,686	uknown
How it was built:	automatically	automatically
Update policy:	no updates	no updates
Access:	WWW	WWW, XML-RPC Web service
Ranking of long forms:	based on confidence	based on quality score
Domain information:	only support for finding abbreviation expansions containing particular keyword	no, but domain coverage is narrow
Language of long forms:	English	English
Language information:	-	-
Licensed:	no	no

Table 4. Request parameters for *Abbreviations.com* Web service API.

Parameter	**Value**	**Default value**	**Description**
tokenid	string (required)		Valid developer token id.
term	string (required)		The term a developer would like to search for.
categoryid	integer	*all*	The category to search in.
sortby	character	*p* (popularity)	The order in which the results will be returned. Use one of the following values: *p* for popularity, *a* for alphabetically or *c* for category.
searchtype	character	*e* (exact match)	The search type to perform. Use one of the following values: *e* for exact match, *b* for begins with and *r* for reverse lookup.

Table 5. Response elements for *Abbreviation.com* Web service API.

Field	**Description**
results	Contains all of the query responses.
result	Contains each individual response.
term	The term this result is referring to.
definition	The definition (long form) that was found for this term.
category	The category that this definition belongs to.

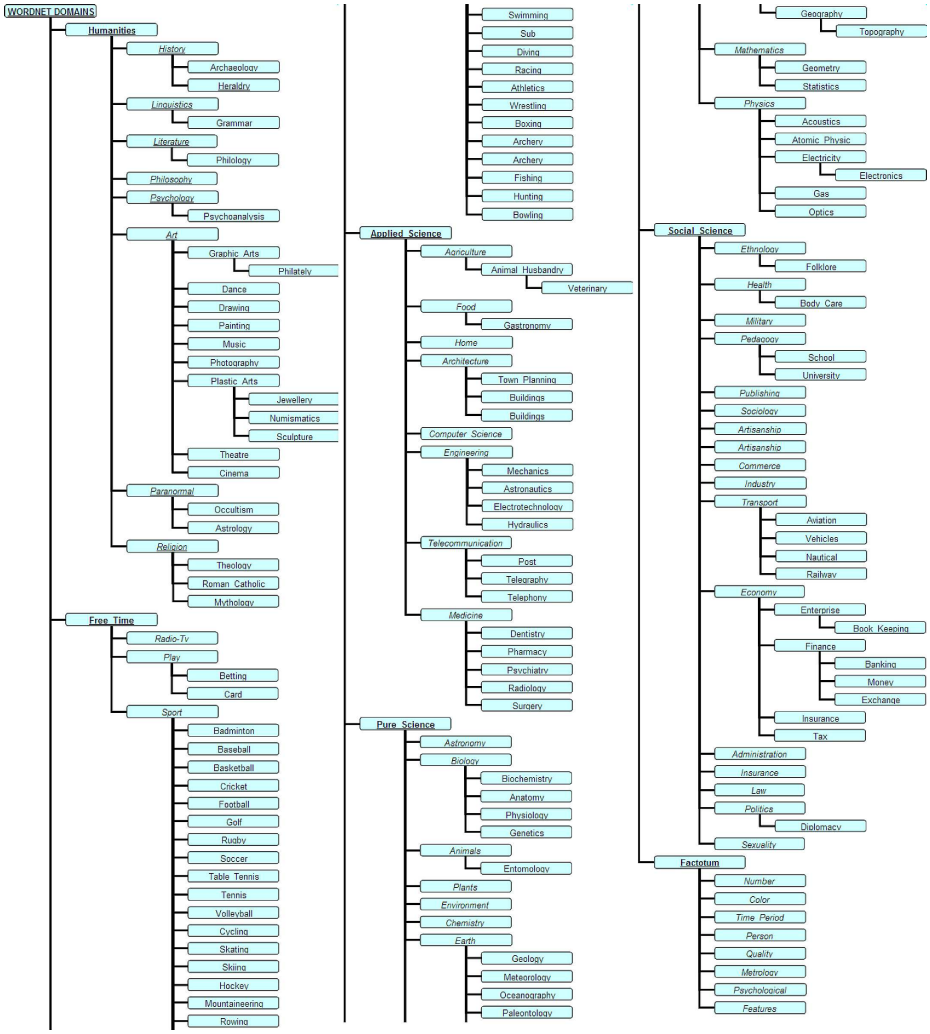

Fig. 1. Domains in WordNet Domains 3.1beta (to be checked).

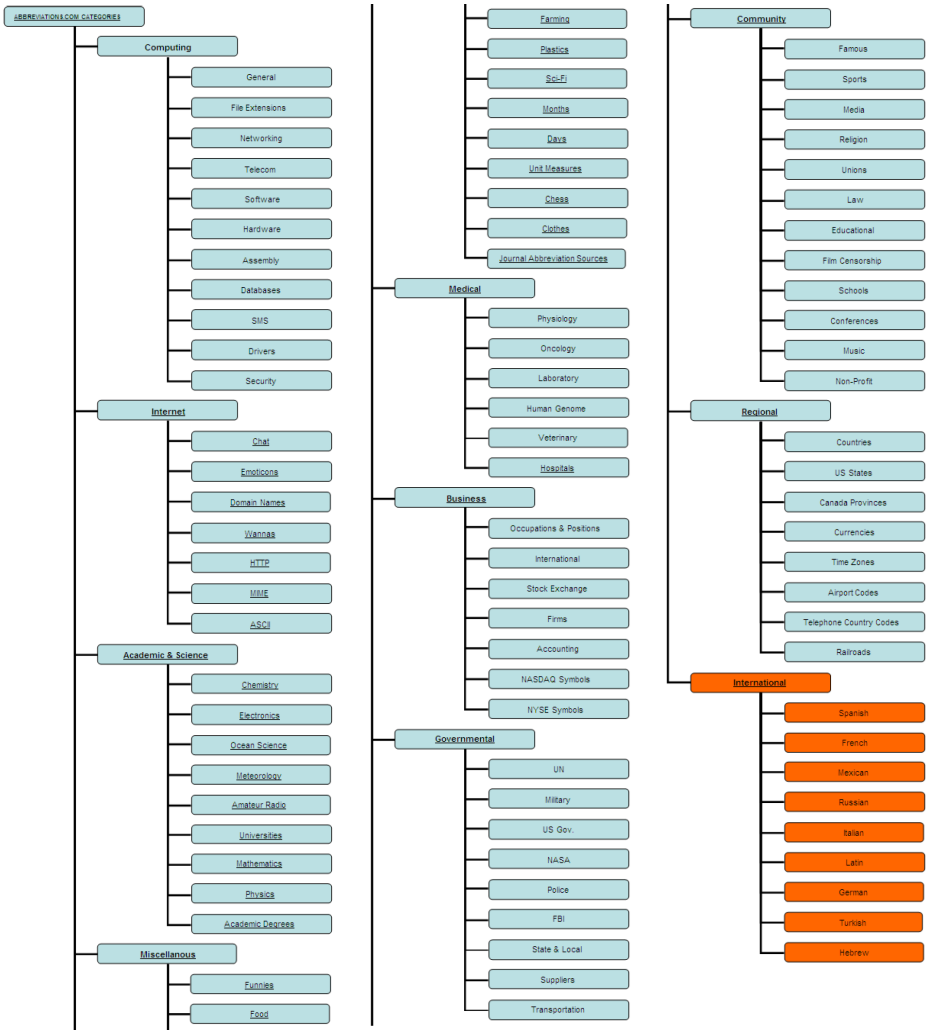

Fig. 2. Categories in Abbreviations.com online dictionary (incomplete).